Computing identity co-reference across drug discovery datasets

Computing Identity Co-Reference Across Drug Discovery Datasets

Christian Y A Brenninkmeijer[1], Ian Dunlop[1], Carole Goble[1], Alasdair J G Gray[2], Steve Pettifer[1], and Robert Stevens[1]

[1] School of Computer Science, University of Manchester, UK.
[2] Department of Computer Science, Heriot-Watt University, UK.

Abstract. This paper presents the rules used within the Open PHACTS (`http://www.openphacts.org`) Identity Management Service to compute co-reference chains across multiple datasets. The web of (linked) data has encouraged a proliferation of identifiers for the concepts captured in datasets; with each dataset using their own identifier. A key data integration challenge is linking the co-referent identifiers, i.e. identifying and linking the equivalent concept in every dataset. Exacerbating this challenge, the datasets model the data differently, so when is one representation truly the same as another? Finally, different users have their own task and domain specific notions of equivalence that are driven by their operational knowledge. Consumers of the data need to be able to choose the notion of operational equivalence to be applied for the context of their application. We highlight the challenges of automatically computing co-reference and the need for capturing the context of the equivalence. This context is then used to control the co-reference computation. Ultimately, the context will enable data consumers to decide which co-references to include in their applications.

1 Introduction

Within the life sciences there has been a proliferation of databases published, with the 2013 NAR database issue listing over 1,500 [8]. An increasing number of these are being published as linked data – either directly, e.g. the recent publication of RDF data by the EBI[3] [17] or through projects such as Bio2RDF [5] – forming a web of linked data. A key data integration challenge is identifying the "same" concept across these datasets.

While there have been attempts to provide global identifiers for a concept, e.g. with life sciences identifiers [7], these have not gained widespread use. Consequently, there is no global, or life sciences, identifier scheme used by all datasets to identify a given concept; each dataset uses their own identifier scheme, leading to a proliferation of identifiers for (notionally) the same concept [12].

Then there is the problem of what does it mean to be the *same*; just representing the truth is not a tenable position. As demonstrated by Halpin *et al* [15]

[3] `http://www.ebi.ac.uk/rdf/` accessed November 2013

there are many interpretations of the `owl:sameAs` relationships that exist in the data. Life sciences datasets contain complimentary and overlapping data, modelled with different levels of granularity depending upon the purpose of the data capture. Therefore, when we refer to the data being about the "same" concept, we are not looking for identical representations but rather saying that these two complimentary records can be treated as being *operationally equivalent*. However, the notion of operational equivalence depends upon the use to which the data will be put and thus can only be decided by the user or the application they use.

Linked data allows a user to navigate their way through the web of data by following links from one resource to a related resource. Consequently not every pair of datasets are linked since you can navigate your way through the web of data. However there are many scenarios where you need to know all of the equivalent URIs. For example, to power a linked data integration platform such as the drug discovery platform being developed by the Open PHACTS project [14]. Co-reference services such as sameas.org [11] and BridgeDb [16] provide a look-up service for discovering equivalent identifiers. However, they are a melting pot of equivalence as they do not consider the context of why two things are equivalent. For example, a search on sameas.org with the URI for the UniProt record for *"Insulin Receptor (homo-sapien)"*[4] results in 30,866 equivalent URIs; the first 1,292 of which are for DBPedia gene entries. Clearly these cannot all be equivalent, particularly since the UniProt record is a protein and not a gene.

In order for scientists to trust mappings, they need to understand the context of the equivalence claim and who is making the claim. By providing the context – in scientific terms – together with the provenance of the mapping – how it was made and by whom – the scientist can understand the notion of equivalence captured and make an informed decision about whether to include it in their application.

In this paper we

- Identify the challenges of identity co-reference across datasets (Section 2);
- Discuss the metadata required to describe a dataset and capture the context of its links to other datasets (Section 3);
- Present the rules used to control co-reference computation and their usage in the Open PHACTS Identity Management Service (IMS) (Section 4).

2 Multiple Identifiers, but are they the same?

Information relevant for drug discovery research is sourced from a variety of overlapping datasets. For example, information about drugs can be retrieved from DrugBank [20], while data about the chemical compounds that compose the drug are available from ChEMBL [10], ChemSpider [21] and DrugBank, and details of the target – typically a protein – that the drug interacts with are available from ChEMBL and UniProt [22]. Since each of these datasets is

[4] `http://www.uniprot.org/uniprot/P06213` accessed September 2013

modelled with a different focus, and have their own identifier scheme, when can we say that two records are truely equivalent? In some cases it is straightforward. An entry in ChemSpider and an entry in ChEMBL that share the same InChI will report about the same chemical, e.g. *"imatinib mesylate"*. However, when we consider the drug entry in DrugBank, e.g. *"Gleevec"*, there can be multiple InChI entries associated, e.g. the *"gleevec"* entry contains the InChI for both *"imatinib"* and *"imatinib mesylate"*. In this case, are the records the "same". For a scientist interested in *"gleevec"* they would be, but for someone interested only in *"imatinib mesylate"* perhaps not.

Many datasets contain links to other related datasets. For example, UniProt includes links to several related datasets. However the nature of these links are not captured; in the case of the RDF export of UniProt they are all stated as `rdfs:seeAlso`. It is therefore hard to automatically reuse such links due to the differing natures of the datasets and meaning of the link. A case in point would be the relationships stated between UniProt – a protein sequence dataset – and Protein Data Bank (PDB) [2] – a 3-dimensional protein structure dataset. Due to the differences in the representations and the data gathering techniques the concepts in these datasets are not, in the strictest sense, equivalent, i.e. there is not a 1:1 isomorphism between the data instances. In particular, the UniProt record for the insulin receptor protein (P06213[5]) links to 18 PDB entries. These in turn map back to six UniProt entries; one of which is the insulin receptor protein we started with.

For users and applications to trust and reuse equivalence relationships, they need to understand what notion of equivalence is being claimed. UniProt, in their RDF export, weaken their links to `rdfs:seeAlso` to avoid making inaccurate claims, but this reduces the knowledge conveyed. At the other extreme, the datasets in the Linked Data Cloud tend to be very relaxed about their claims of "equivalence" and widely use, or misuse, the predicate `owl:sameAs`; typically they do not intend the strict semantics of `owl:sameAs`. As such, these links need to be used with caution. Such context will enable applications to choose which links to include. For example, for the vast majority of drug discovery research – almost all – it would be acceptable to use "equivalence" links between gene entries in Entrez Gene and protein entries in UniProt as genes are often used as proxies for the protein that they encode. For example, when searching for data about a target, it is common to enter the gene name as the search term. However, there are those who would require that such links are not included, e.g. working in very niche specialisms or on edge cases.

3 Describing Datasets and their Links

For effective linking between datasets, and to enable trust in their use in applications, it is essential to understand what has been linked and how. This requires descriptions of the datasets and the links themselves.

[5] `http://www.uniprot.org/uniprot/P06213` accessed Sept. 2013

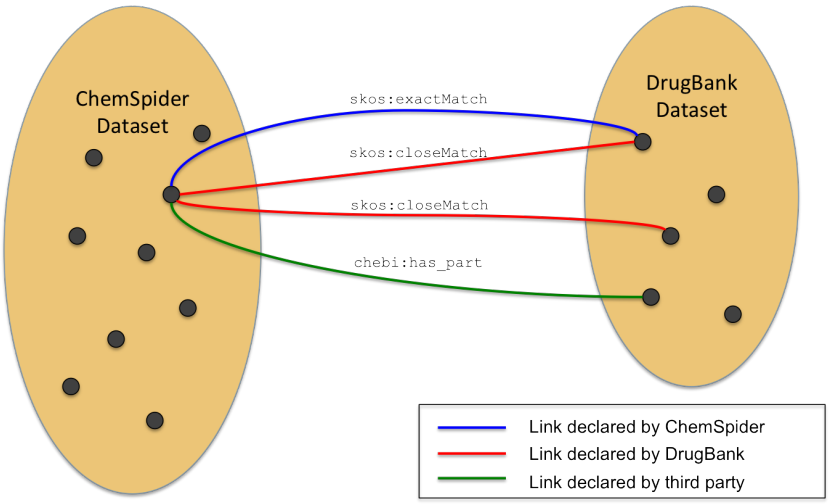


Fig. 1: Illustrating example links for a ChemSpider entry from three linksets connecting ChemSpider with DrugBank. Each linkset has a separate publisher using a different link predicate relationship.

A dataset description is essential for data discovery and to enable consumers to use the data. It is a means to provide core metadata about a dataset, e.g. its title, description, and license information. It can also convey information about how to access the dataset, e.g. a SPARQL endpoint location, how the data is modelled, i.e. which vocabularies have been used, and key statistics about the data, e.g. number of triples, number of subjects, etc.

The Vocabulary of Interlinked Datasets (VoID) [1] provides a vocabulary of terms and a deployment model for dataset descriptions. Where possible, VoID recommends re-using Dublin Core terms, e.g. for providing the title (`dct:title`) and license (`dct:license`). VoID itself provides predicates for expressing access and statistical information. A key feature of VoID is the ability to embed the dataset description with the data. This is achieved by the data linking back to its description using the `void:inDataset` predicate. VoID also introduces the notion of a *Linkset* which is a collection of links between a pair of datasets. The linkset description captures the context of the links, i.e. which datasets are linked using what predicate. Some benefits of providing separate linksets are that the linksets can develop independently of the datasets and even be provided by third-parties, as well as being used in co-reference services. Figure 1 illustrates four example links for a ChemSpider entry to the DrugBank dataset. These links are drawn from three distinct linksets published by different providers using a variety of link relationships.

However, VoID does not prescribe which properties must be provided and those that are more optional. This makes the general re-use of VoID dataset descriptions difficult as there is no guarantee that the information you need will

be provided. For example, for the pharmaceutical companies involved with Open PHACTS it is essential to understand the licensing restrictions of the dataset, but this information may not be present in the dataset description. There is also no notion of capturing the version of a dataset, which is essential to know when linking between datasets. To overcome these challenges in the Open PHACTS project, we have defined a checklist of properties that must be provided [13]. We have also identified additional vocabulary terms for capturing the context of the linkset, e.g. the Provenance, Authoring and Version vocabulary (PAV) [6] is used to provide the version number of a dataset.

4 Identity Co-Reference Computation

For systems such as the Open PHACTS Discovery Platform (OPSDP) [14], links are required between several datasets. However, it is not practical to require that each pair of datasets is directly related. As such we propose that co-reference identities can be transitively computed from those that are supplied. However, care needs to be taken to avoid computing inaccurate co-references as may result from a chain of links with varying meaning. In this section we detail some alternative strategies, the problems with them, and outline the approach currently adopted by the Open PHACTS Identity Mapping Service (IMS) [4].

4.1 Link Predicate

A VoID linkset includes three key pieces of information: the dataset that is the subject of the link triples, the dataset that is the object of the link triples, and the predicate used in the links. Based on this information it is feasible to compute transitive co-reference links based on the properties of the link predicate. For example, given the linksets $A \xrightarrow[p]{} B$ and $B \xrightarrow[p]{} C$ which link the datasets A and B, and B and C respectively, with the link predicate p, and let p be the predicate `owl:sameAs`, then it follows through the properties of the link predicate that we have the linkset $A \xrightarrow[p]{} C$ which links datasets A and C.

However, as shown by Halpin *et al* [15] when `owl:sameAs` links are used they are often not truely equivalent. Additionally, as indicated in Section 2, many different link predicates are in use in life sciences datasets. These predicates each have different properties, e.g. `rdfs:seeAlso` is neither transitive nor symmetric. As such, it is not possible to compute a complete network of co-reference identifiers across the set of required datasets based on OWL reasoning over the link predicates. Therefore we need a custom approach to computing the transitive co-reference across datasets that requires more than just the link predicate as a means of control.

4.2 Linkset Justification

The limitation of the linkset predicate approach stems from the generality of the linking predicate and thus the lack of domain knowledge that it conveys. One

Term	Justification
Chemical entity `sio:SIO_010004`	The concepts linked represent the same chemical entity.
Gene `sio:SIO_010035`	The concepts linked are conceptually the same gene.
InChI Key `cheminf:CHEMINF_000059`	The concepts linked have the same InChI Key.
Protein `sio:SIO_010043`	The concepts linked are conceptually the same protein.
Protein coding gene `sio:SIO_000985`	A gene resource and a protein resource are being treated as being equivalent

Table 1: A subset of the vocabulary terms used to capture the justification of a linkset and the operational equivalence that is interpreted. `sio` represents the Semantic Science Integrated Ontology namespace and `cheminf` the Chemical Information Ontology namespace.

approach to overcome this, whilst still retaining the notion of a VoID linkset, would be to mint a new linking predicate for each notion of equivalence; these could be created as sub-properties of existing mapping predicates. However there is a major social barrier to such an approach; gaining consensus on the required linking predicates and updating the existing links in the datasets to use these new link predicates. As such, it is unlikely to gain traction.

Another alternative is to annotate the linkset descriptions with additional contextual data; this enables the use of the existing links unchanged. We term this the *justification* for the linkset; the notion captured is the scientific interpretation of the operational equivalence applied by the linkset. For example, two chemical datasets, A and B, that are linked because they have the same InChI string would express this relationship in the linkset VoID header with the triples

```
:A-B_Linkset void:linkPredicate skos:exactMatch ;
             bdb:linksetJustification cheminf:CHEMINF_000059 .
```

where `:A-B_Linkset` is the resource that describes the linkset, the link predicate is declared to be `skos:exactMatch`, and the justification is specified using the BridgeDb vocabulary namespace (`bdb`[6]) with the value taken from the Chemical Information Ontology namespace (`cheminf`[7]). The linkset can be expressed as $A \xrightarrow[p]{j} B$ where the justification j is `cheminf:CHEMINF_000059` and the link predicate p is `skos:exactMatch`. The set of supported justifications within the Open PHACTS IMS can be found in [13]; a subset of which are included in Table 1. A key advantage of this approach is that it extends rather then changes the existing data.

Based on the justification of linksets, we can compute transitive linksets. For example, we can generate a linkset between datasets A and C through some

[6] `http://vocabularies.bridgedb.org/ops` to appear soon

[7] `http://semanticscience.org/resource/CHEMINF_000059` accessed Sept. 2013

intermediary dataset B if there is a linkset between A and B and one between B and C such that both linksets have the same justification. Definition 1 formally gives the rule for computing transitive linksets based on their linkset justification. Note that we do not require that the linksets have the same link predicate. The resulting transitive linkset is given the weaker of the two link predicates with a hierarchy of

$$\texttt{owl:sameAs} \preceq \texttt{skos:exactMatch} \preceq \texttt{skos:closeMatch} \preceq \texttt{rdfs:seeAlso}.$$

Thus, if p was the link predicate `owl:sameAs` and r the link predicate `rdfs:seeAlso`, the computed linkset $A \xrightarrow[r]{j} C$ would have the link predicate `rdfs:seeAlso`.

Definition 1 (Transitive computation based on linkset justification). *Given datasets A, B, and C, linksets $A \xrightarrow[p]{j} B$ and $B \xrightarrow[r]{j} C$ both with the justification j and link predicates p and r respectively then we can generate the linkset*

- $A \xrightarrow[r]{j} C$ *if* $p \preceq r$;
- $A \xrightarrow[p]{j} C$ *if* $r \prec p$.

By iteratively applying the rule given in Definition 1 it is possible to compute chains of linksets that use the same justification. However it is possible to enter an infinite cycle; thus the IMS implementation prevents the same linkset being used more than once in a chain. As part of the provenance of the computed linkset, the linksets that are used to compute it are tracked.

4.3 Permitting Cross-type Equivalence

Within the life sciences it is common to use gene names as proxies for protein names since gene names are shorter and more standardised. It is easy for a human to distinguish when this is being done but impossible for a computer to distinguish. A key requirement for the OPSDP is to permit a user to enter with a gene name that is then resolved to a URI for that gene. However, it should be possible to retrieve information about the target – a protein, or group of proteins – for which the entered gene name is a proxy. This means that it must be possible to state that a gene and a protein are *operationally equivalent*, i.e. to have equivalence across semantic types.

This is straightforward using the linkset justification approach. We introduce a new justification for a protein coding gene (`sio:SIO_000985`), see Table 1. The complication comes when computing the transitive linksets to enable a user entered gene to relate to the protein information in each of the datasets. The transitive computation now needs to support equivalence across semantic types with different justifications. In particular we want to support chains of one or more protein linksets, a protein-gene linkset, and one or more gene linksets. It is important to prevent the use of protein-gene linksets to go from a gene to a

protein and back again; this is to prevent a chain of links whereby we end up with protein X being claimed to be the same as protein Y due to crossing the semantic type boundary multiple times.

Additional contextual information is required from the linkset description. Specifically, the semantic type of the data being linked. This needs to be captured at the linkset level since datasets can contain multiple semantic types, e.g. ChEMBL, DrugBank and Ensembl [9]. Two additional predicates are used to capture the semantic type: `bdb:subjectsType` and `bdb:objectsType`. Note that these mirror the VoID predicates for specifying the datasets that are linked.

Definition 2 extends the co-reference transitive computation rule given in Definition 1 to support cross-type mappings. Note that for simplicity we have omitted the link predicate from the rules in Definition 2. These are derived in the same way as for the rule given in Definition 1. The first clause of Definition 2 is a combination of the clauses in Definition 1, with the additional constraint that all of the datasets involved are of the same semantic type. The second clause allows for a linkset of the same semantic type and a linkset with a cross semantic type justification to be combined, with the resulting linkset being given the cross-type justification.

Definition 2 (Cross-type transitive computation). *Let A, B, and C be datasets, τ and τ' be semantic types with $\tau \neq \tau'$, and j and j' be linkset justifications such that j links datasets with the same type and j' links datasets across types. Then the following two rules hold for transitive linkset computation:*

- *Same semantic type and justification*

$$\frac{A_\tau \xrightarrow{j} B_\tau \ \wedge \ B_\tau \xrightarrow{j} C_\tau}{A_\tau \xrightarrow{j} C_\tau}$$

- *Cross semantic type and justification*

$$\frac{\left(A_\tau \xrightarrow{j} B_\tau \ \wedge \ B_\tau \xrightarrow{j'} C_{\tau'}\right) \ \vee \ \left(A_\tau \xrightarrow{j'} B_{\tau'} \ \wedge \ B_{\tau'} \xrightarrow{j} C_{\tau'}\right)}{A_\tau \xrightarrow{j'} C_{\tau'}}$$

By iteratively applying Definition 2 it is possible to compute the co-reference of URIs across proteins and genes. Note that the definition only permits a single cross-type link justification, e.g. "protein coding gene", to be used in any chain, although arbitrary numbers of same type justifications, e.g. "same protein" or "same gene", can be applied. This is due to the consequence of the second rule being given the cross-type link justification. Thus, a chain of links resulting in inaccurate mappings is prevented.

5 Open PHACTS IMS Implementation

The Open PHACTS IMS implementation is an extended version of BridgeDb to support cross-references over linked data sources, i.e. supporting the use

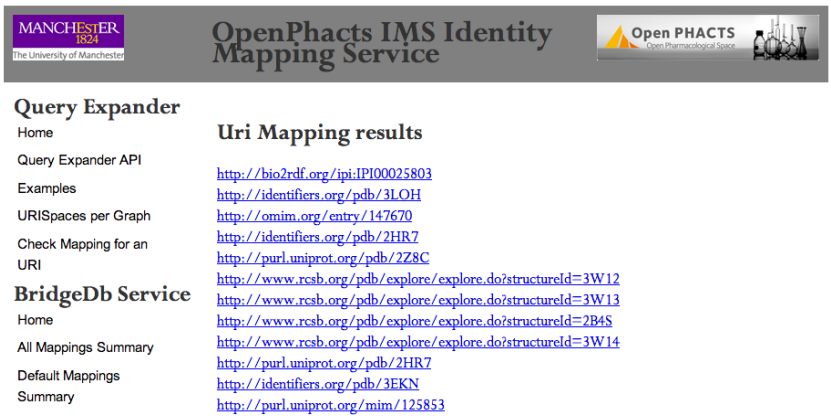


Fig. 2: Screenshot of the Open PHACTS Identity Mapping Service web interface showing the results for the UniProt entry for insulin receptor.

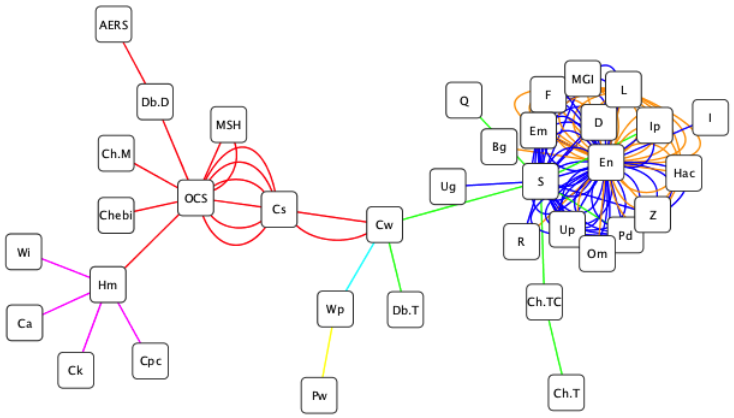
of URIs to represent records in datasets. The source code is available from `https://github.com/openphacts/IdentityMappingService` and the service is accessible through the Open PHACTS API, `https://dev.openphacts.org/`.

The IMS has implemented the transitive co-reference computation rule given in Definition 2. Computed linksets are created with dataset descriptions giving full details of the datasets linked and the justification for the link. A screenshot of the web interface to the IMS is shown in Figure 2 with the results for a look-up for the UniProt insulin receptor URI.

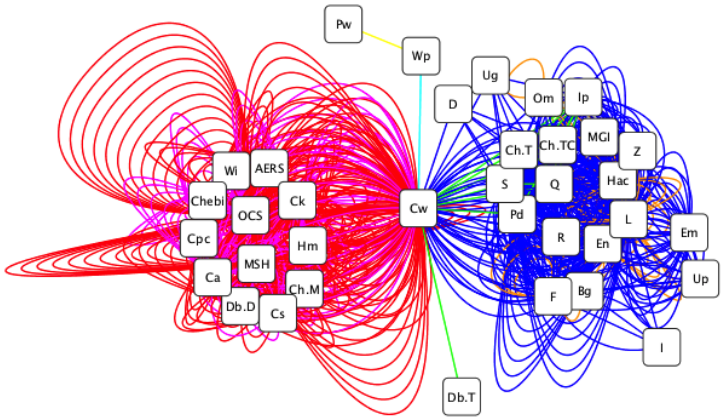
5.1 Result of Co-reference Computation

For the 1.3 release of the OPSDP, the IMS was supplied with 104 linksets from 9 providers linking 37 datasets and containing 7,096,712 links. These are shown in the visualisation in Figure 3a; nodes represent datasets and edges the linksets between them, with the colour signifying the linkset justification. Note that the large number of linksets is a consequence of splitting links based on their semantic type – resulting in multiple linksets between some datasets – and in the case of Ensembl the linksets were further split by species with 12 species covered. The visualisation highlights the Open PHACTS design decision to use a small number of datasets as mapping centres; chemical alignment is performed through the Open PHACTS Chemical Registration Service [19] – OCS in the figure – with a few through the Human Metabolome Database (HMDB) [23] – Hm in the figure – proteins centre around UniProt – S in the figure – and genes around Ensembl – En in the figure.

Following the transitive co-reference computation, there are 883 linksets containing a total of 17,383,846 links. These are shown in the visualisation in Figure 3b. The left side of the visualisation is dominated by the linksets that match InChIs via the Open PHACTS Chemical Registration Service while the right side



(a) Visualisation showing connectivity of the linksets provided as input to the IMS.



(b) Visualisation showing connectivity of the linksets after co-reference computation.

Legends

Colour	Justification	URI
Red	InChI	`cheminf:CHEMINF_000059`
Purple	Chemical entity	`sio:SIO_010004`
Green	Protein	`sio:SIO_010043`
Orange	Gene	`sio:SIO_010035`
Blue	Protein coding gene	`sio:SIO_000985`
Light blue	Pathway	`sio:SIO_001107`
Yellow	Pathway name	`edam:data_2342`

Code	Dataset
AERS	Adverse events reporting system
Bg	BioGrid
Ca	Chemical abstracts service
Chebi	Chemical entities of biological interest
Ch.M	ChEMBL molecule
Ch.T	ChEMBL target
Ch.TC	ChEMBL target components
Ck	KEGG Compound
Cw	ConceptWiki
Cpc	PubChem Compound
Cs	ChemSpider
D	Saccharomyces genome database
Db.D	Drugbank drugs
Db.T	Drugbank targets
Em	European nucleotide archive
En	Ensembl
F	FlyBase
Hac	HGNC accession number

Code	Dataset
Hm	Human metabolite database
I	InterPro
Ip	International protein index
L	NCBI Gene
MGI	Mouse genome informatics
MSH	Medical subject headings
Pw	Pathway ontology
Om	Online mendelian inheritance in man
OCS	OPS Chemical registration service
Pd	Protein databank
Q	NCBI Reference Sequence Database
R	Rat genome database
S	UniProt
Ug	UniGene
Up	UniParc
Wi	Wikipedia
Wp	Wikipathways
Z	Zebrafish information network

Fig. 3: Visualisations showing connectivity of the linksets.

is dominated by protein coding genes due to the application of the cross-type co-reference rule. The visualisation highlights a second Open PHACTS design decision; to use ConceptWiki[8] – Cw in the figure – as the source for text-to-URI translation. This is shown by its connectivity to every other dataset; one of the reasons for computing the co-references. As a consequence of these co-reference computations, users of the OPSDP are able to enter with any chemical, protein, or gene URI known to the IMS and retrieve information about chemicals or targets.

5.2 Evaluation of Co-reference Computation

There are several benefits to the co-reference computation approach that has been implemented as part of the OPSDP IMS; not least of which is the increased inter-connectivity across the datasets.

First it eases the burden on the dataset providers; with only a small number of additional metadata triples being required to provide justifications and semantic types. This allows systems such as the OPSDP to exploit the fact that many datasets already contain links within their data to instances in other datasets; particularly those who publish their data as RDF. However it cannot be expected that they can link to every dataset that is required for every possible use case. Thus, by applying the co-reference computation we are able to infer additional connectivity across the data.

Second the co-reference computation tightly controls what can be equated, e.g. only chemical entities can be related through an InChI, and which justifications are allowed to cross semantic types, e.g. protein coding gene relating genes and proteins. These safeguards ensure that the result of the co-reference computation matches the expectations of the domain scientists.

User evaluation of the results of the co-reference computation is on-going with domain scientists. The IMS has been successfully deployed by the OPSDP enabling the integration of data across the data sources shown in Figure 3.

6 Related Work

Identifiers.org [18] provide a linked data identifier for many life sciences datasets. This consists of a URI constructed according to the rules of the identifier scheme of the underlying dataset. However, they do not attempt to identify co-referent identifiers. The Identifiers.org approach is complimentary to the co-reference work reported here. The IMS accepts and returns the Identifiers.org form of URI for each of the datasets.

Bio2RDF [5] is another closely related approach. Bio2RDF republishes existing datasets as RDF where the source data has been originally published as database dumps or in other formats. Where the original datasets contain links to other datasets these are published with the RDF. These links could be used

[8] `http://ops.conceptwiki.org/` accessed September 2013

as a source of mapping information for the IMS and the IMS already returns the Bio2RDF identifier as an alternative URI.

BridgeDb [16] and sameas.org [11] are co-reference services similar to the IMS. The goal of sameas.org is to ingest the links of as many linked data sources as possible as such it has a much broader coverage of topics than the IMS. It provides an API for returning all known equivalent URIs. However, there is no curation of the links nor context for the data. Additionally, to the best of our knowledge, there is no transitive computation across the co-referent URIs. BridgeDb is a life sciences focused database identity cross-reference service. However, it does not track the equivalent URIs for the database entries. It also does not characterise the database cross-reference nor their context. The IMS is an extension of BridgeDb to provide a URI look-up service that understands the context of the links as well as the linking relationship.

Another closely related area of research is focused on generating tools for identifying links between pairs of datasets; either at the schema level *Ontology Matching* or at the instance level. The latter of these is most relevant to the work in this paper. Since 2009 there has been an instance matching track[9] in the annual ontology matching competition[10] to compare such tools. The most recent results are available from `http://www.instancematching.org/oaei/imei2013/results.html`. The links generated by these instance matching tools could be used as input to the IMS.

7 Conclusions

In this paper we presented the challenges for co-reference across life sciences datasets that stem from each dataset using their own identifier scheme. We have argued that there is not a one size fits all notion of equivalence across concepts in these different datasets since they model the data at different levels of granularity, e.g. should a drug entry be equated to an entry about the chemical compound. Additionally, users of the data want to apply varying notions of equivalence based on the task they are performing, e.g. should genes and proteins always be equivalent. As such we have proposed that the notion of operational equivalence should be captured in the linksets that relate a pair of datasets as the *justification* for the linkset. The advantage of stating it as a justification rather than a mapping predicate is that existing linksets can be easily extended.

Due to the fact that each pair of datasets is not related in the web of data, we have developed rules for transitively computing co-reference across datasets. To support scenarios where genes and proteins should be equated, the co-reference computation allows the crossing of semantic types. We presented a rule for preventing undesired co-references being computed whilst ensuring that concepts within a given type are completely covered. This has been implemented in the

[9] `http://www.instancematching.org/oaei/imei2013/results.html` accessed November 2013

[10] `http://oaei.ontologymatching.org/` accessed November 2013

Open PHACTS Identity Mapping Service. Details of how the IMS is used within the OPSDP can be found in [14,4].

As future work we will allow applications to apply different *scientific lenses* [3] over the co-reference network to vary the notion of operational equivalence being applied, i.e. to activate different combinations of linksets based on their justification. These lenses depend upon the justifications used to compute the co-references.

Acknowledgements

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement number 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007- 2013) and EFPIA companies' in kind contribution.

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary. Note, W3C (Mar 2011), http://www.w3.org/TR/void/
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic acids research 28(1), 235–42 (Jan 2000), http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102472&tool=pmcentrez&rendertype=abstract
3. Brenninkmeijer, C.Y.A., Evelo, C., Goble, C., Gray, A.J.G., Groth, P., Pettifer, S., Stevens, R., Williams, A.J., Willighagen, E.L.: Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision. In: Proceedings of 2nd International Workshop on Linked Science 2012 (LISC2012) Colocated 11th International Semantic Web Conference 2012. CEUR-WS.org, Boston, MA, USA (2012), http://ceur-ws.org/Vol-951/paper5.pdf
4. Brenninkmeijer, C.Y.A., Goble, C., Gray, A.J.G., Groth, P., Loizou, A., Pettifer, S.: Including Co-referent URIs in a SPARQL Query. In: 4th International Workshop on Consuming Linked Data. Sydney, Australia (Jul 2013)
5. Callahan, A., Cruz-toledo, J., Ansell, P., Dumontier, M.: Bio2RDF Release 2 : Improved Coverage , Interoperability. In: ESWC 2013. pp. 200–212. Springer, Montpellier, France (2013)
6. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J.G., Goble, C., Clark, T.: PAV ontology: Provenance, Authoring and Versioning. arXiv.org (Apr 2013), http://arxiv.org/abs/1304.7224, submitted to Journal of Biomedical Semantics
7. Clark, T., Martin, S., Liefeld, T.: Globally distributed object identification for biological knowledgebases. Briefings in bioinformatics 5(1), 59–70 (Mar 2004), http://www.ncbi.nlm.nih.gov/pubmed/15153306
8. Fernández-Suárez, X.M., Galperin, M.Y.: The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. Nucleic acids research 41(Database issue), D1–7 (Jan 2013), http://nar.oxfordjournals.org/content/early/2012/11/30/nar.gks1297

9. Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J.P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., Searle, S.M.J.: Ensembl 2013. Nucleic acids research 41(Database issue), D48–55 (Jan 2013), http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531136&tool=pmcentrez&rendertype=abstract
10. Gaulton, A., Bellis, L.J., Bento, a.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P.: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research 40(Database issue), D1100–7 (Jan 2012), http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245175&tool=pmcentrez&rendertype=abstract
11. Glaser, H., Jaffri, A., Millard, I.: Managing Co-reference on the Semantic Web. In: WWW2009 Workshop: Linked Data on the Web (LDOW2009). Madrid, Spain (Apr 2009)
12. Goble, C., Stevens, R.: State of the nation in data integration for bioinformatics. Journal of biomedical informatics 41(5), 687–93 (Oct 2008), http://www.ncbi.nlm.nih.gov/pubmed/18358788
13. Gray, A.J.G.: Dataset descriptions for the Open Pharmacological Space. Working draft, Open PHACTS (Sep 2013), http://www.openphacts.org/specs/datadesc
14. Gray, A.J.G., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C.Y.A., Burger, K., Chichester, C., Evelo, C.T., Goble, C.A., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., Williams, A.J.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semantic Web (2014), http://iospress.metapress.com/index/J3J12776V103821U.pdf
15. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In: International Semantic Web Conference (1). LNCS, vol. 6496, pp. 305–320. Springer, Shanghai, China (Nov 2010)
16. van Iersel, M.P., Pico, A.R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B.R., Evelo, C.T.: The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics 11(5) (Jan 2010), http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2824678&tool=pmcentrez&rendertype=abstract
17. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S., Martin, M., Le Novere, N., Parkinson, H., Birney, E., Jenkinson, A.: The EBI RDF platform: Linked open data for the life sciences. Bioinformatics Application Note (2014), accepted for publication November 2013
18. Juty, N., Le Novère, N., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic acids research 40(Database issue), D580–6 (Jan 2012), http://nar.oxfordjournals.org/content/40/D1/D580
19. Karapetyan, K., Tkachenko, V., Batchelor, C., Sharpe, D., Williams, A.J.: Rsc chemical validation and standardization platform: A potential path to quality-

conscious databases. In: 245th American Chemical Society National Meeting and Exposition. New Orleans, LA, USA (April 2013)
20. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., Wishart, D.S.: DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research 39(Database issue), D1035–41 (Jan 2011), http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013709&tool=pmcentrez&rendertype=abstract
21. Pence, H.E., Williams, A.J.: ChemSpider: an online chemical information resource. Journal of Chemical Education 87(11), 10–11 (2010), http://pubs.acs.org/doi/abs/10.1021/ed100697w
22. The UniProt Consortium: Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic acids research 41(Database issue), D43–7 (Jan 2013), http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531094&tool=pmcentrez&rendertype=abstract
23. Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., Macinnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J., Querengesser, L.: HMDB: the Human Metabolome Database. Nucleic acids research 35(Database issue), D521–6 (Jan 2007), http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1899095&tool=pmcentrez&rendertype=abstract